8-17-2018

Epigenetic Profiling of Active Enhancers in Mouse Retinal Ganglion Cells

Jian Xing
jian.xing@uconn.edu

Epigenetic Profiling of Active Enhancers in Mouse Retinal Ganglion Cells

Jian Xing

B.S., University of Science and Technology Beijing 2016

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

At the

University of Connecticut

2018

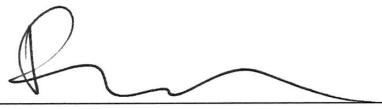
APPROVAL PAGE

Masters of Science Thesis

Epigenetic Profiling of Active Enhancers in Mouse Retinal Ganglion Cells

Presented by

Jian Xing, B.S.

Major Advisor 
Feliks (Ephraim) Trakhtenberg

Associate Advisor 
Royce Mohan

Associate Advisor 
Paola Bargagna-Mohan

University of Connecticut

2018

ACKNOWLEDGEMENTS

The author thanks Drs. Feliks (Ephraim) Trakhtenberg

, Royce Mohan, Paola Bargagna-Mohan, and for their assistance revising this thesis

Table of Contents

List of Figures

Abstract

Retinal ganglion cells (RGCs) are projection neurons of the eye, which process and pass visual information collected in the eyes to the brain. However, epigenetic regulation of RGC fate specification remains poorly understood, in large part due to the technical challenges associated with purifying RGCs, which comprise only 1% of all retinal cells, and performing ChIP-seq profiling on a small number of cells. To overcome these limitations, we have purified RGCs from multiple mice by immunoppanning for Thy1, a surface marker on RGCs, and analyzed pulled chromatin by ChIP-seq for histone 3 acetylated at K27 (H3K27ac), which is an epigenetic marker of active (i.e. accessible) enhancers. We also utilized recently developed ChIP-seq Kit for processing small cell number samples and ultra low chromatin input (truChIP Ultra Low Cell Chromatin Shearing Kit, Covaris; Zymo-Spin ChIP Kit, Zymo Research; and Accel-NGS 2S Plus DNA Library Kit, Swift Biosciences). Immunoprecipitation chromatin was sequenced and mapped to mouse reference genome (mm10), peaks were called using MACS2 (sheared RGC chromatin was used for background control), and peak annotation was performed using the HOMER software. We have identified enhancers that are active in RGCs and cross-referenced them to the mouse cell types, and some were novel and unique to RGCs. Next, using HOMER software, we performed DNA motif analysis in the identified enhancer regions and predicted the transcription factors to those we have identified as expressed in RGCs by mRNA-seq. These analyses enabled us to associate the transcription factors that are expressed in RGCs with the enhancers that are active in RGCs, and also to predict which of these transcription factors could bind to the same enhancers, and thus may cooperate with each other in regulating gene expression in RGCs. Taken together, our epigenetic profiling of RGCs predicted active enhancers, which are potentially unique to RGCs, as well as predicted the interactions of transcription factors and enhancers, which are involved in determining RGC-specific gene program.

Chapter 1: Introduction/Background

1.1 Overview

Retina ganglion cells (RGCs) are projection neurons of the eye, which located near the ganglion cell layer of the retina. They integrate the visual information collected in eyes and pass the information through the optic nerve into the brain for further processing.

In gene control, regulation of transcription is the most common form. Enhancers located upstream or downstream from an associated gene are DNA regions that transcription factors can bind to and form a complex that recruits transcriptional machinery of the promoter of the gene to enhance the transcription. By binding to a specific DNA sequence, the transcription factors can control the rate of transcription of genetic information from DNA to mRNA. The activity of transcription factors enables genes to be specifically regulated during developing in variant types of cells.

DNA sequencing is a laboratory approach to determine the precise sequence of nucleotides of a DNA molecule. The method was developed by Frederick Sanger. Next generation sequencing technologies that allow millions of DNA molecules to be sequenced at a time have become essential in studies of genomics, epigenomics, and transcriptomics. These recent technologies enable us to sequence DNA and RNA much more quickly and cheaply than the previous approach -- Sanger sequencing. Chromatin immunoprecipitation followed by high throughput sequencing, which also known as ChIP-seq, is popular in sequencing techniques to study genome-wide protein-DNA interactions. The approach of ChIP-seq combines chromatin immuniprecipitation with DNA sequencing to identify the binding sites of DNA associated proteins.

Our goal is to integrate epigenetic and transcriptional regulation of RGCs development, in order to advance the understanding of RGCs biology and identify therapeutic targets that could help improve health and survival of injured RGCs. We identified the active enhancers during RGCs differentiation and maturation.

The epigenetic studies can not only help us to investigate epigenetics of RGCs development but also predict which of the transcription factors could cooperate together in recruiting the transcriptional machinery to promoters of genes in RGCs.

For my thesis project, I focused on performing epigenetic profiling of active enhancers in mouse retinal ganglion cells to study their development. We purified retinal ganglion cells from many mice by immunopanning for the surface marker Thy1 on them. The mice are come from different conditions include embryonic day 18, postnatal day 0 and adult. Each condition has two replicates with five mice ten eyes. The we analyzed chromatin by ChIP-seq for epigenetic markers of active enhancers. After sequencing, I mapped the DNA reads to the mouse reference genome so that I could identify the peaks that represent active enhancers in retinal ganglion cells. For the control sample, I used fragmented background DNA without purifying. After getting peaks alignment to the mouse genome and cross-referencing to the mouse enhancers identified in multiple tissues from the ENCODE database, I discovered a number of novel enhancers that maybe specific to only retinal ganglion cells because we used highly purified retinal ganglion cells which only represent 1% of all retinal cells. Then I integrated the ChIP-seq result with the gene expression profile of retinal ganglion cells that I have performed using our mRNA-seq dataset. By integrating them together, I could find the first down stream gene of each enhancer that expressed in retinal ganglion cells. I hypothesized that the enhancer is involved in regulating the down stream gene expression. I also associated the transcription factors that are expressed in retinal ganglion cells with the active enhancers and predicted which of these transcription factors could bind to the same enhancers. They were close enough to cooperate with each other in forming transcriptional complexes that regulate gene expression in retinal ganglion cells. I validated the novel enhancers by qPCR.

This chapter provides background into the relevant experimental materials, sample types and bioinformatics tools studied in my M.S. thesis research.

1.2 Retinal Ganglion Cells

1.2.1 Retina

Vision is a complicated process that requires components of eye and brain to work together. The first step in the process of the sense is carried out in the retina. The retina is a complex transparent light-sensitive tissue that lines the inner surface of the back of the eyeball consisting of several layers. Only one of the layer in the mature mammalian retinas contains light-sensitive photoreceptor cells. The photoreceptor cells have two types – rods and cones, which are differentiated structurally by their distinctive shapes and functionally by their sensitivity to different kinds of lights[1]. Light must pass through the overlying layers to reach them. Rod photoreceptors can detect motion and provide black- white vision which means they function well in low-light level. Cones are more prominent in animals that are active during the day (most of the mammals) because they are responsible for central vision, color vision and perform better in medium and bright light. Rod cells are located throughout the retina and cones are concentrated in a small central area of the retina – macula. These two types of cells can take light focused by the cornea, lens and convert it into chemical and nervous signals which are transported to visual centers in the brain through the optic nerves[2].

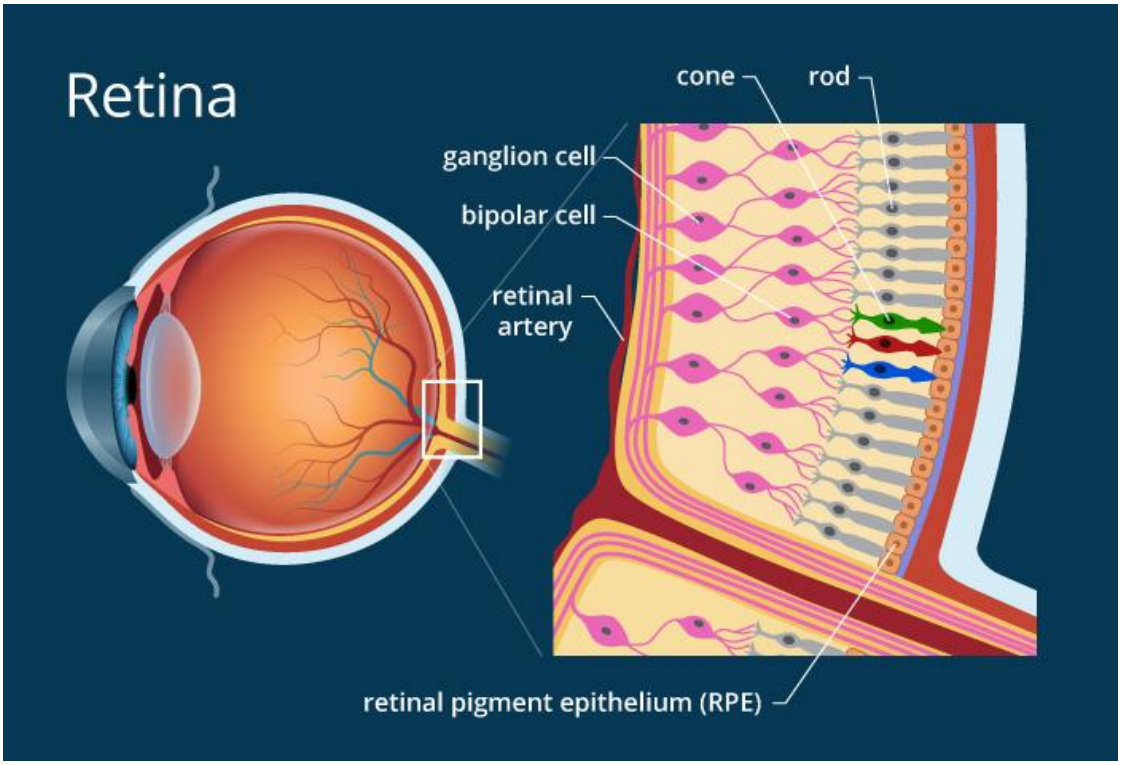


Figure 1: The structure of retina. Light are focused by the cornea and lens into the retina, where vision begins (Gary Heiting, OD).

1.2.2 Ganglion cells

Retinal neuronal neurons have six types: bipolar cells, ganglion cells, horizontal cells, retinal amacrine cells, rod and cone photoreceptors. Retina ganglion cells are a type of projection neuron which are the final output of the retina. They collect visual information from retinal neurons – photoreceptors via intermediate neuron types: bipolar and amacrine cells, and convey information to the rest of the brain for further processing[3]. The visual information is a type of chemical messages that is sensed by receptors on the ganglion cell membrane and the chemical messages are transformed into intracellular electrical signals by transmembrane receptors[4]. These messages are integrated within the dendrites and cell bodies of ganglion cells and transmitted through axon of ganglion cells. The optic nerve collects all the axons of the ganglion cells so that axons can terminate in brain

visual centers. Retinal ganglion cells only represent one percent of all retinal cells but they are all photosensitive. Their axons form the retinohypothalamic track and conduce to circadian rhythms and pupillary light reflex.

1.3 Cis-regulatory Element

Cis-regulatory elements (CREs) are regulatory elements of non-coding DNA that contain binding sites for transcription factors and play essential roles in cell development since they control development and physiology by regulating gene expression. The regulatory events contribute to the gene expression programs that determine the state of cells and the potential for differentiation into new cell types[5].

Promoters and enhancers are two types of cis-regulatory elements. DNA sequences with promoter activity initiate transcription at transcription start sites and sequences with enhancer activity activate transcription initiation at promoters[6]. Regulation of transcription is the most common form of gene control, and the activity of transcription factors enables genes to be specifically regulated during development in variant types of cells. Cis-regulatory elements typically regulate gene transcription by binding to transcription factors.

1.3.1 Promoter

Promoters are regions of DNA where transcription of a gene controlled by RNA polymerase begins. Promoters are always about 100-1000 base pairs long and located directly upstream on the DNA or near the 5' end of transcription start sites of genes. RNA polymerase and some necessary transcription factors can bind to the promoter sequences and initiate transcription (Figure 2). There are two types of promoters: the first one is core promoter elements that general transcription machinery including the RNA polymerase core enzyme and other general transcription factors could bind to. The recruitment of RNA polymerase II (Pol II) to the core promoter region can help the occurrence of the initiation protein-coding genes and

distinct non-coding regions transcription[7]. Another type is regulatory elements that bind sequence-specific transcription factors that include transcriptional activators and repressors[5].

1.3.2 Enhancer

Enhancers are DNA short non-coding regions (50-1500bp) that transcription factors could bind to and form a complex that recruits transcriptional machinery to the promoter of a gene to regulate expression. When bound by specific transcription factors, enhancer sequences can enhance the transcription of an associated gene to increase the likelihood that transcription of the gene will occur.

The region of enhancer sequences can be located up to 1 Mbp away from the gene, upstream or downstream from the transcription start site and also within introns so that they generally function independently at various distances from their target promoters[8]. Enhancer sequences function at a distance by forming chromatin loops so that enhancer and target gene can be brought into proximity. Enhancer elements are associated with histone modifications and histone H3K27ac can distinguish active enhancers from inactive or poised enhancer elements that contain H3K4mel alone[9]. Active enhancers interact with transcription factors to regulate specific gene expression show more type-specificity of cell than promoters[6].

1.3.3 Transcription factor

Transcription is the key process to transcribe the genetic information from DNA to make a protein where the DNA sequence of a gene is copied into a RNA molecule. Transcription factors (TFs) are proteins that regulate the transcription of genes by binding to a specific DNA sequence. These proteins help determine which genes are active in each cell, in other words, determine whether the gene's DNA is transcribed into RNA. The chemical reactions of RNA synthesis are catalyzed by the enzyme RNA polymerase, using the gene's DNA as a template. In this process, transcription factors can control when, where and how efficiently the RNA polymerase perform.

A single transcription factor can bind to many cis-regulatory elements, and hence control the expression of several genes, while many transcription factors may act together to regulate the expression of one gene[10]. Transcription factors are a very diverse family of proteins and generally function in protein complexes which means a complex of transcription factors can cooperate with each other to regulate the expression of a specific gene. They may bind directly to the promoter regions of DNA (Figure 2), which locate at the upstream of a gene, or directly to the RNA polymerase molecule. In general, transcription factors work alone or with others to form a complex. They not only can act as a activator but also can act as a repressor to promote or block the recruitment of RNA polymerase to specific genes.

Transcription factors contain at least one DNA-binding domain (DBD), with which they can attach to a specific sequence of DNA adjacent to the genes that they regulate. DBD is the unique characteristic of transcription factors, since many other proteins that also help to regulate the gene expression such as histone acetyltransferases, histone deacetylases and coactivators are lack DNA-binding domains.

Theoretically, there are two classes of transcription factors: general transcription factors and specific transcription factors. General transcription factors are ubiquitous and interact with the core promoter region around the transcription start sites. Specific transcription factors always bind upstream of the initiation site and they vary considerably depending on the recognition sequences.

There are three main mechanisms that transcription factors are used. The first one is stabilizing or blocking the binding of the RNA polymerase to DNA molecule. The second one is to catalyze the acetylation or deacetylation of histone proteins to weaken or strengthen the association of DNA with histones so that they make the DNA more or less accessible to transcription, thereby up-regulating or down-regulating transcription. The third one is to recruit coactivator or corepressor proteins to the transcription factor DNA complex.

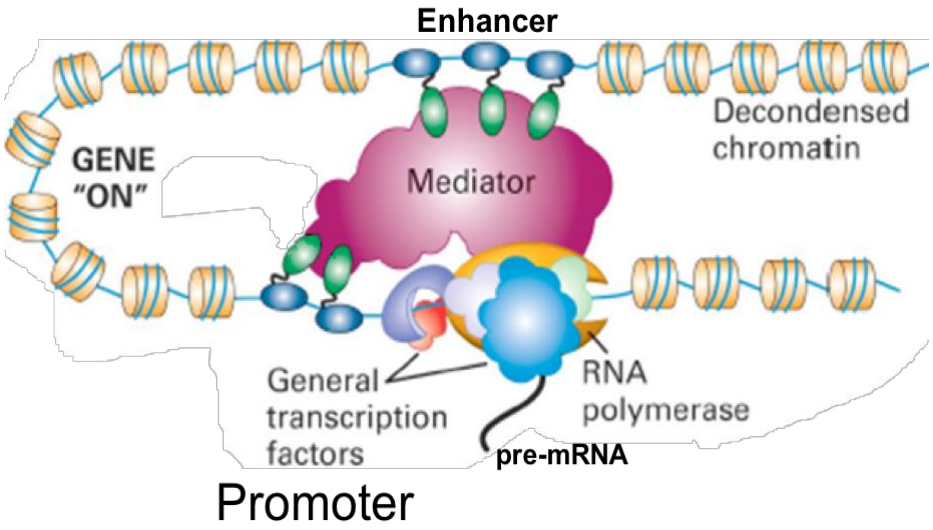


Figure 2: The transcription factors bind to DNA regions and form a complex which recruits transcriptional machinery to the promoter of a gene to regulate its expression.

1.4 Sequencing

In genetics, sequencing is the general term that refers to any techniques designed to determine the precise order of nucleotides in a nucleic and acid molecule.

1.4.1 Sanger sequencing

Sanger sequencing is the first-generation sequencing method that is based on the selective incorporation of chain-terming dideoxynucleotides by DNA polymerase in vitro DNA replication. Sanger sequencing, which was first developed by Fred Sanger and his colleagues in 1977, has been the most widely used sequencing method for forty years. Sanger method can be used to determine the sequences of small fragments extracted from DNA molecules. The fragments are mapped to assemble the sequence of reference genome based on the overlapping regions. Although genomes are now sequenced by other alternative methods that are faster and cheaper, Sanger sequencing is still in a wide use for the sequencing

of individual fragments of DNA that are generated through polymerase chain reaction[11].

1.4.2 Next generation sequencing

Next-generation sequencing, also known as high-throughput DNA sequencing, is a method that enables rapid sequencing of the base pairs in DNA or RNA molecules. By using next-generation sequencing, millions or billions of DNA strands can be sequenced in parallel, yielding redundant throughput and minimizing the need for the fragment-cloning method[10]. Next-generation sequencing can be used for transcriptomic analysis of mRNAs, small RNAs, noncoding RNAs, genome-wide methylation assays and high-throughput chromatin immunoprecipitation assays.

1.4.2.1 DNA sequencing

DNA sequencing is the approach to determine the precise order of nucleotides of a DNA molecule. Sequencing a short piece of DNA is relatively straightforward while sequencing an entire genome requires breaking the DNA of genome into shorter reads, sequencing the reads and assembling the small reads into a single long consensus. With the development of sequencing technology, genome sequencing is much faster and cheaper than it was when using the first generation sequencing method.

Illumina sequencing is one of the common DNA sequencing method that determine the series of base pairs in DNA molecules with high data accuracy, simple workflow and a broad range of applications. This innovative and flexible sequencing system enables a wide use in genomics, transcriptomics and epigenetics. The Illumina sequencing workflow is composed of 4 basic steps – sample preparation, cluster generation, sequencing and data analysis. The preparation adds adapters to the ends of the DNA and through reduced cycle amplification, additional motifs are introduced, such as sequencing binding site, induces and regions complementary to the flow cell oligos. In the clustering process, each short fragment is isothermally amplified. The flow cell is a glass

slide with lanes, which are the channels coated with a lawn, composed of two types of oligos. Hybridization is enabled by one the two types of oligos on the surface. This oligo is complementary to the adapter region on one of the fragment strands. A polymerase creates a complement of the hybridized fragment. The double stranded molecule is denatured and the original template is washed away. Then the strands are amplified through bridge amplification. In the amplification process, the strand folds over and the adapter region hybridizes to the second type of the oligo on the flow cell. Polymerases generate the complimentary strand forming a double stranded bridge. The bridge is then denatured, resulting in two single stranded copies of the molecule that are tethered to the flow cell. The amplification is repeated over and over again to occur simultaneously for millions of clusters resulting in clonal amplification of all the fragments. After bridge amplification, the reverse strands are cleaved and washed off, leaving only the forward strands. The 3' prime ends are blocked to prevent unwanted priming. Sequencing begins with the extension of the first sequencing primer to produce the first read. With each cycle, fluorescently tagged nucleotides compete for addition to the growing chain. Only one is incorporated based on the sequence of the template. After the addiction of each nucleotide the clusters are excited by a light source and a characteristic fluorescent signal is emitted. This process is so-called sequencing-by-synthesis. The emission wave length, along with the signal intensity, determines the base call. For a given cluster, all identical strands are read simultaneously and hundreds of millions of clusters are sequenced in a massively parallel process. After the completion of addition of the first read, the read product is washed away, the index one read primer is then introduced and hybridized to the template. The read is generated as the first read. After completion of the index read, the read product is washed off again and the 3' prime ends of the template are deprotected. The template then folds over and binds to the second oligo on the flow cell. Index two is read in the same manner as index one. Polymerases extend the second flow cell oligo forming a double stranded bridge. This double stranded DNA is linearized and the 3' prime ends are blocked. The original forward strand is cleaved off and washed away only leaving the reverse strand. Read two begins to be sequenced with the introduction of the read two sequencing primer. The sequencing steps are repeated until the desired read length

is achieved and then the read two product is washed away. During the laser excitation, the image is captured and the identity of the second base is recorded to determine the base call (Figure 3). This entire process generates millions of reads, representing all the fragments from the target DNA molecules. All of the sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time[12]. The Illumina provides two fastq files containing the raw sequence data of both ends. The data are aligned and compared to a reference genome and sequencing differences are identified.

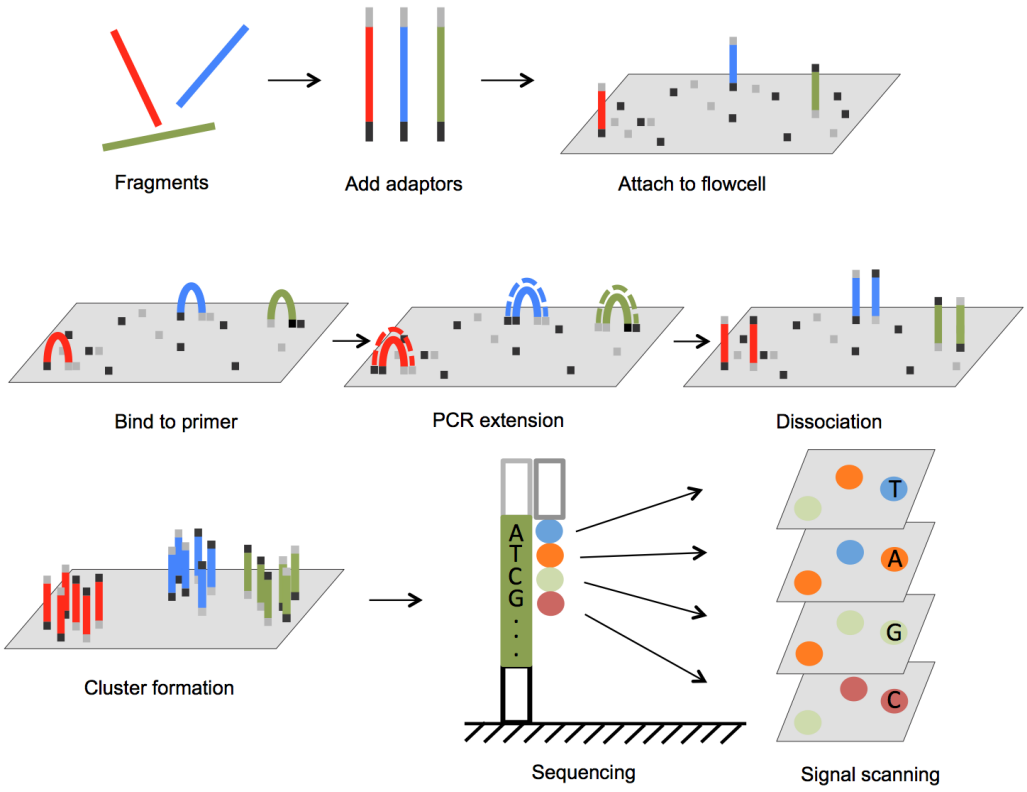


Figure 3: The process of Illumina sequencing (Ben Faircloth)

1.4.2.2 RNA sequencing

RNA sequencing uses next-generation to transcriptome profiling. The transcriptome is the complete set of transcripts in a cell that represents only small percentage if the genetic code can be transcribed into RNA molecules including non-coding sequences. Each gene may produce more than one variant of mRNA since

there are alternative splicing, RNA editing, even alternative initiation and termination sites so that the transcriptome can capture a level of more complex genetic information than genome sequence[13]. We can use RNA sequencing to analyze the cellular transcriptome in order to understand the development of cells. RNA sequencing can only determine the sequences that are actively expressed in the cells, while DNA sequencing can give a genetic profile of entire genome. RNA sequencing also provides a more precise measurement to identify transcripts and their isoforms[14].

In general, total or fractionated (such as poly(A)+ selection) RNA molecules are converted to a library of cDNA fragments that are attached with adapters to one or both ends. After amplification, each molecule is sequenced in a high-throughput manner to get short sequences from one end or both ends(Figure 4)[14 15]. Next-generation sequencing technologies, such as Illumina sequencing, SOLiD sequencing and Roche 454 sequencing, can be used for RNA sequencing. After short reads having been obtained, the resulting reads are either mapped to a reference genome, reference transcripts or assembled *de novo* without the genomic sequence. The reads can be classified into three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a genome scale transcription map that contains both the transcriptional structure or level of gene expression.

RNA sequencing is often used to perform transcriptome profiling, the identification of novel transcripts, expressed SNPs, alternative splicing, and detection of gene fusion[15].

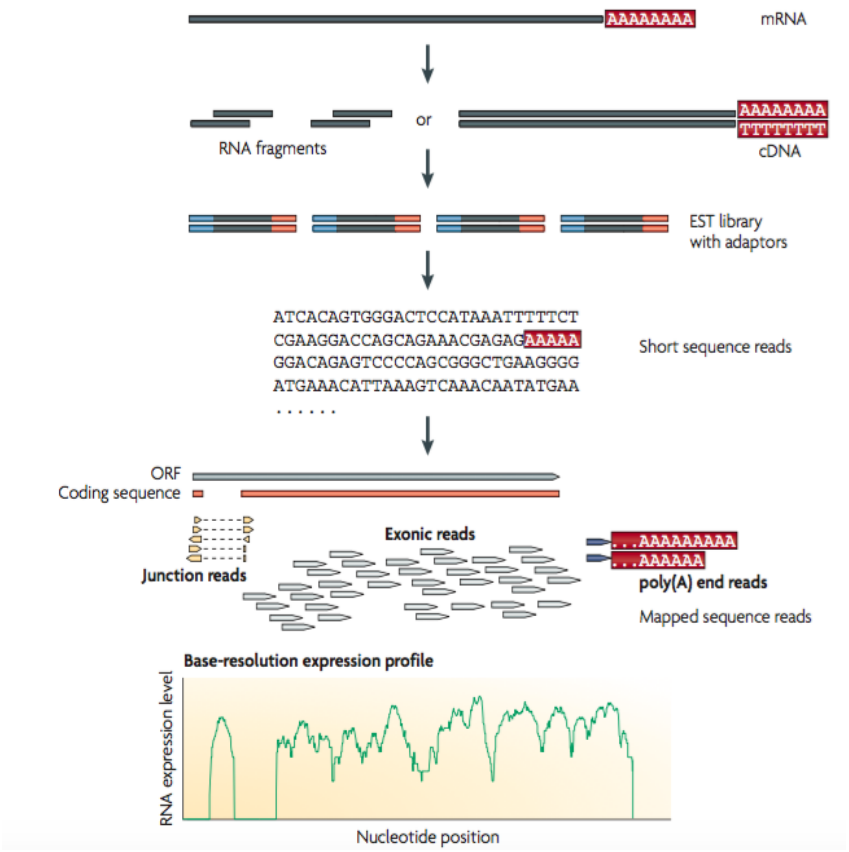


Figure 4: mRNA sequencing overview (Nature Review Genetics, 2018)

1.4.2.3 ChIP sequencing

Chromatin immunoprecipitation followed by sequencing is popular in sequencing techniques to study genome-wide protein-DNA interactions. The approach of ChIP sequencing combine chromatin immunoprecipitation with DNA sequencing to identify the binding sites of DNA associated proteins[16].

In a ChIP-seq workflow, DNA sequences that has been bound to protein by crosslinking is sheared and the bound DNA reads are pulled by immunoprecipitation. After been purified, the reads are converted a library of cDNA fragments for sequencing (Figure 5). The results will provide the binding sites for the proteins of interest.

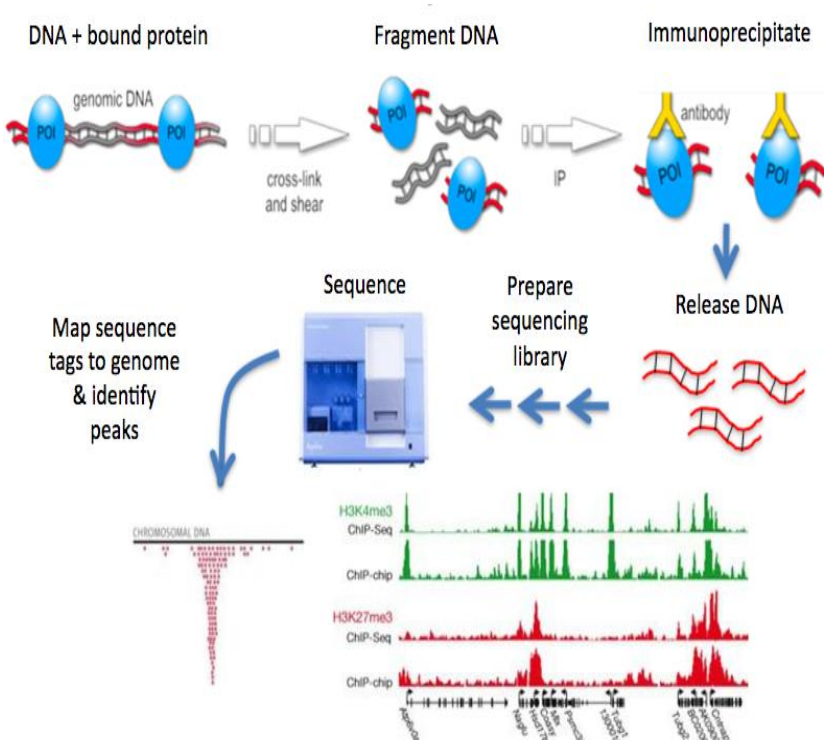


Figure 5 ChIP sequencing overview (Stuart M. Brown)

1.5 Bioinformatics analysis

Despite there are many benefits of next-generation sequencing, the data generated are challenging to work with. The next-generation sequencing has a much higher error rate than Sanger sequencing. Furthermore, the next-generation sequencing only produces short fragments, ranging from 100-300 nucleotides in length as well as the datasets are very large, always larger than 100 gigabytes[17]. The next-generation sequencing experiments must be analyzed by a robust, efficient and statistically principles pipeline to get significant results. A number of experimental and bioinformatics innovations are explored to help us address these challenges. Bioinformatics is an interdisciplinary field that uses computational approaches including collection, storage, retrieval, manipulation and modelling for data analysis, visualization and prediction through the development of algorithms and software to extract knowledge from experimental data. Bioinformatics analysis is wide used in high-throughput data-generating experiments, including genomic sequence determinations and

measurements of gene expression. The bioinformatics analysis allows us to determine whether the data are consistent with the hypothesis.

Chapter 2: Materials and Methods

2.1 Cell purification

Retinal ganglion cells (RGCs) were purified from embryonic day 18, postnatal day 0, 5 and adult mice eyes single cell suspension by immunopanning for Thy1 (CD90, MCA02R, Serotec) after depletion of macrophages (using anti-mouse macrophage antibody, AIA31240, Accurate Chemical) and washing off the non-adherent cells, following an established protocol[18]. There are two replicates of each age and five mice's eyes were collected in each replicate.

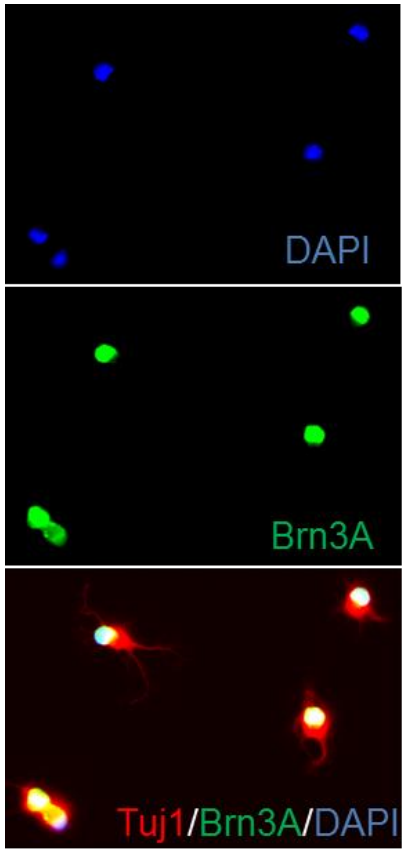


Figure 6 Representative image of P5 RGCs immunostained for RGC marker Brn3A and neuronal marker Tuj1, at 12 hours in culture after immunopanning (Scale bar, 100 μm).

2.2 RNAseq library

Before sequencing, highly abundant ribosomal RNAs must be removed from the sample so that mRNA can be detected to perform transcriptome profiling[15]. Polyadenylation is the addition of a polyA tail to a messenger RNA. In eukaryotes, polyadenlation is a part of the process that produces mature mRNA for translation. Since most messenger RNAs contain a polyA tail while structural RNAs do not have them, we used polyA selected RNA method in our experiment. The polyA RNA selection technique enables the rapid and highly specific enrichment of polyadenylated RNAs from total RNA samples. The embryonic day 18 and postnatal day 5 sample libraries that were prepared from polyA-selected RNA sequenced 100bp from each ends on HiSeq 2000 Sequencer (Illumina)[19 20]. All samples from different ages included two biological replicates for raw reads.

2.3 ChIPseq library

In ChIP-seq analysis, the small cell number populations of purified retinal ganglion cells result in low put ChIP-seq reads. We used the kits which improve DNA sequencing result by covering more of the genome can overcome this limitation. The histone 3 acetylated at K27 (H3K27ac) is an epigenetic marker of active enhancers while the histone 3 acetylated at K4 trimethylation can mark active promoters which absent in the H3K27ac DNA regions. We used H3K4me3 to identify active enhancers and found the regions of active enhancers by H3K27ac in ChIP-seq pulling chromatin. Purified retinal ganglion cells were chemically cross-linked to their bound DNA by formaldehyde treatment. We used the truChIP Ultra Low Cell Chromatin Shearing Kit (Covaris) and M220 Focused-ultrasonicator (Covaris) to isolate and shear the chromatin with the parameters recommended by the manufacturer for small cell number that lead to excellent ChIP-seq performance in other studies[21 22]. After that, we used Zymo-Spin ChIP kit (Zymo), which is used to optimize

for small input amounts of chromatin from mammalian cells. It also has been validated on the material generated using Covaris methods. Immunoprecipitations were performed overnight after clearance using the anti-H3K27ac antibody, which were validated for ChIP-seq in mouse cells and Protein G Magnetic Beads (Pierce). For each age, two biological replicates and a non-ChIP' ed sheared chromatin control were generated. Libraries were prepared from purified ChIP DNA using the Accel-NGS 2S Plus Kit (Swift Biosciences) that utilizes a proprietary adapter attachment chemistry which can decrease the bias and support inputs as low as 10 pg. All samples from different ages included two biological replicates for raw reads.

ChIP-seq samples were sequenced by NextSeq (Illumina) at the depth of 16M PE read. Raw data from the Next-generation sequencing machine often display in FASTQ format, which contain short DNA sequence and quality scores. Quality controlled (QC) can be used before reads mapping to help excluding reads with low quality[16]. We used FastQC to remove adapters and trim reads. FastQC provides a simple way to do quality control check on raw sequence data coming from high throughput sequencing machines so that we can get the information whether the data has any problems before further analysis. The process of quality control can filter out the reads whose Phred quality score < 28 and length < 40. The Phred quality score is a standard that measure the quality of the identification of the bases generated during the DNA sequencing so that we can use it to characterize the quality of DNA sequences.

2.4 RNA-seq bioinformatics pipeline

2.4.1 Reads mapping

A number of computational and statistical tools have been used to analyze high-throughput sequence data. At first, the RNA-seq reads (FASTQ files) that were generated from the HiSeq 2000 Sequencer (Illumina) were trimmed by Trimmomatic. Trimmomatic is a fast and flexible command line tool that can be used to trim and crop FASTQ file generated by the Illumina as well as to remove adapters[23]. With the help of Trimmomatic,

FASTQ files were removed technical sequences and filtered with the quality scores. After trimming, the new FASTQ files were mapped to the mouse reference genome which is from the GENCODE database (Version M13) by Hisat2. Hisat2 is an ultra fast and sensitive mapping tool to align the next-generation sequencing reads to the reference genome, using a hierarchical Graph FM index. The large small GFM indexes can collectively cover the whole genome and the global GFM index can represent a population of whole genome. The output of Hisat2 is BAM file. VM13 mouse reference genome containing the comprehensive gene annotation on the reference chromosomes is used as the reference genome. We get the BAM files from the retinal ganglion cells of embryonic day 18 mouse and postnatal day 5 mouse.

2.4.2 Transcripts assembly

Cufflinks packages contain a set of tools to assembly transcripts and do some differential expression analysis for RNA-seq data. The RNA-seq experiments always produce enormous volumes of raw sequencing reads data. Cufflinks is a software tool that developed with mathematics, statistics and computer science idea together. Cufflinks is used to assemble the reads from the aligned BAM file generated by the Hisats2 into transcripts. Cuffmerge merges two or more transcript assemblies while Cuffdiff is used to take the aligned reads from several conditions to find differentially expressed genes and transcripts using a rigorous statistical analysis (Figure 7)[24].

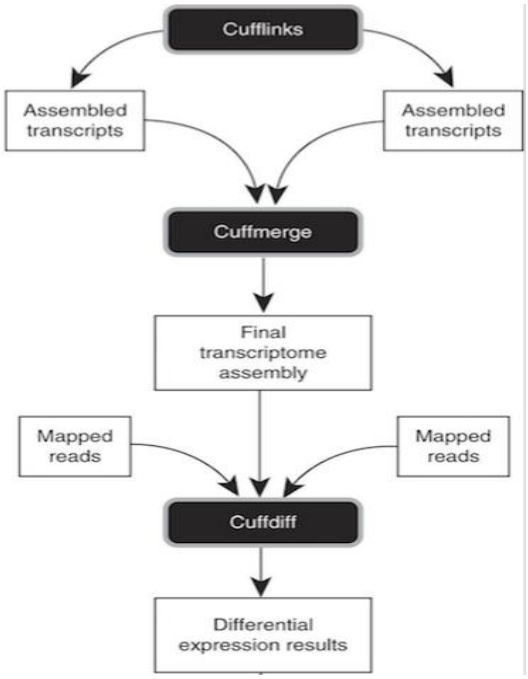


Figure 7 An overview of Cufflinks workflow.

2.5 ChIP-seq bioinformatics pipeline

2.5.1 Reads mapping

The reads that were generated from the sequencer were mapped to the mouse reference genome which is from the GENCODE database (Version M13) by Bowtie2. Bowtie2, which combines the strengths of the full-text minute index with the flexibility and speed of hardware-accelerated dynamic programming algorithms, is an ultrafast and memory-efficient mapping tool for aligning DNA reads to reference genome[25 26]. The index and algorithms can help Bowtie2 to achieve a combination of high speed, sensitivity and accuracy[26]. The output of the Bowtie2 are SAM file, which is sequence alignment map format file. SAM is a text-based format file that can store biological sequences aligned to a reference genome so that it is widely used to store data generated by next generation sequencing technologies.

Since the binary version of SAM files (BAM) are often the input files needed for different analysis program, SAM files are converted to BAM files by SAMtools. SAMtools can provide a set of computational tools that

are used in next-generation sequence analysis to manipulate alignments in the SAM and BAM format. Binary file is an efficient computer file and usually much smaller than a text file that contains an equivalent amount of data. In terms of memory, storing values using numeric formats tends to use less memory so that it is easy to access by computer.

BAM was compressed. After converting, BAM files were sorted by SAMtools because sorting helps to give a better compression ratio. The similar sequences are grouped together while the unmapped reads are put at the end of the file.

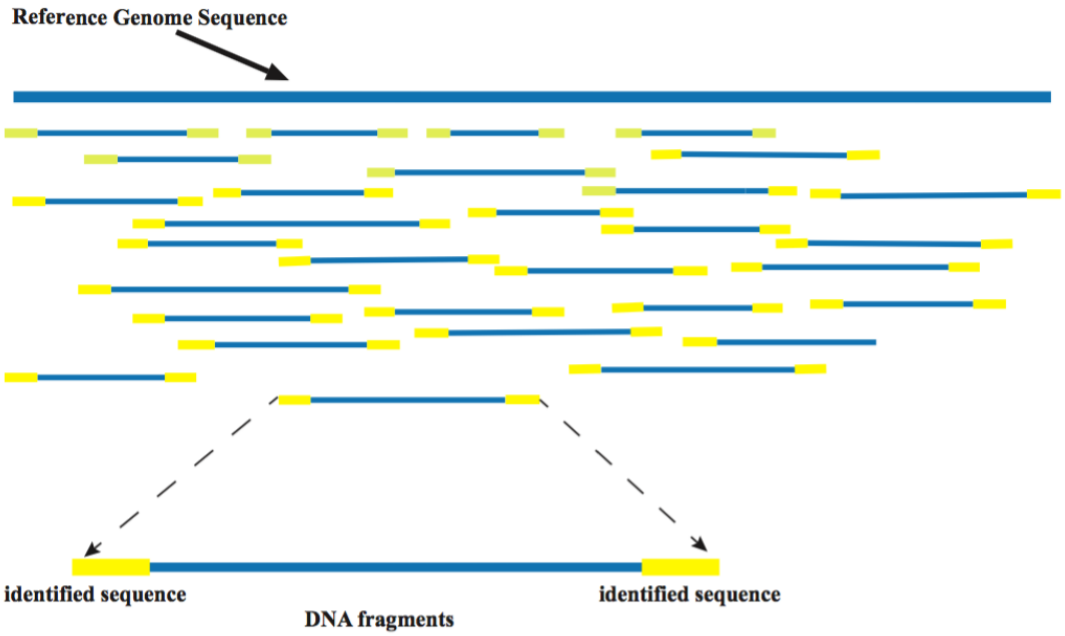


Figure 8: Paired end reads of sequencing data mapped to reference genome

2.5.2 Remove duplicates

The next practice is to remove duplicate reads, namely just keep one read per genomic position, strand and sequencing library[27]. In the DNA sequencing, each nucleotide in the target DNA position is sequenced multiple times to overcome the high error rate. The number of reads that include the target nucleotide is referred to as

coverage. There is the assumption that since sequencing errors are random, if each nucleotide is sequenced multiple times, most reads will give us the correct genetic sequence. Therefore, making a deeper coverage is more reliable to determine the nucleotide in the target position[17]. Polymerase chain reaction (PCR) duplicates are sequence reads that result from multiple sequencing times of the same DNA fragment which may contain erroneous mutation introduced during PCR amplification[17]. PCR duplicates can lead to impediment to the data analysis. The main purpose of removing duplicates in ChIP-seq data is to mitigate the effects of PCR amplification bias introduced during library construction. In addition, after removing duplicates, the data can be gained the computational benefits of reducing the number of reads to be processed in downstream steps. The Picard is a set of command line tools which can manipulate high-throughput sequencing data. When preparing ChIP-seq samples, duplicates can arise during library construction. One of the functions of Picard is to locate and tag duplicate reads in the BAM or SAM files - MarkDuplicates. MarkDuplicates identifies read pairs with the same 5' start position in the alignment process. After duplicates reads are collected, Picard differentiates the primary and duplicates reads by an algorithm that can rank reads by the sums of their quality scores. The BAM files are marked duplicates using Picard to identify duplicate reads. The output of Picard MarkDuplicates was a new BAM file, in which duplicates had been identified and marked by setting the SAM flag 1024 for all but the best read. The best read pair was the one with the highest sum of quality scores. After that BAM files were removed the duplicates that were marked by Picards through SAMtools.

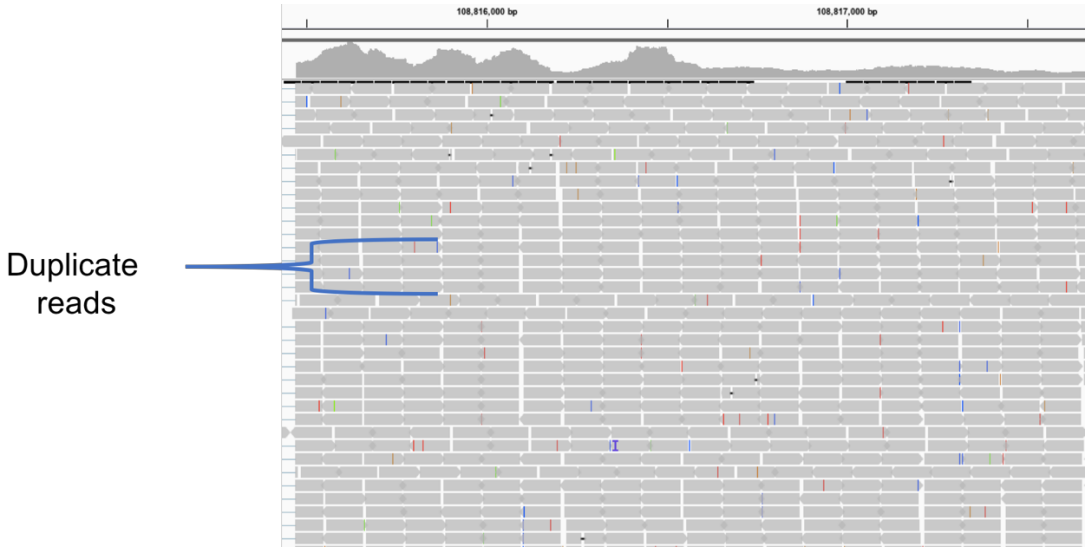


Figure 9 The duplicate reads in raw sequencing data

2.5.3 Peak calling and identification

Peak calling is a computational method used to identify the sites of protein that DNA binding by identifying regions where are enriched with mapping reads. The peak is the region that ChIP-seq reads are enriched (Figure 10).

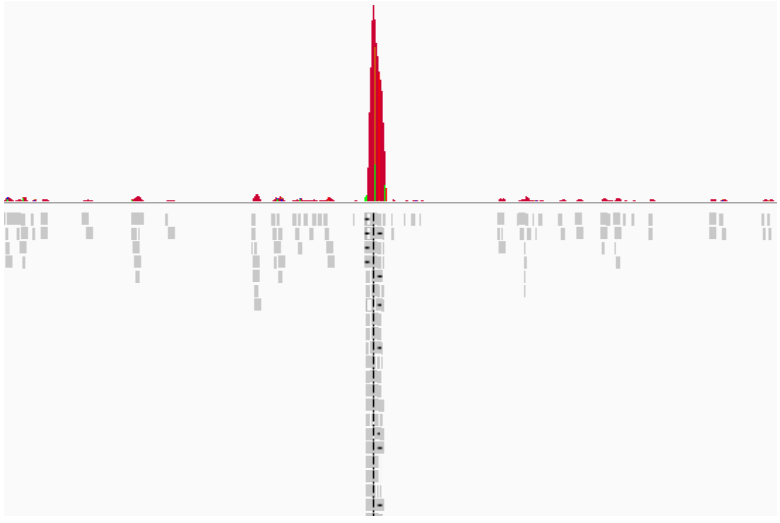


Figure 10: ChIP-seq peak and the coverage of ChIP-seq reads.

After mapping sequence reads to the genome, we used an algorithm, named Model-based Analysis of ChIP-seq (MACS2) to identify transcription factor binding sites. MACS2 is a computational algorithm that identifies genome-wide locations of protein binding sites or histone modification from ChIP-seq data. It captures the influence of genome complexity to evaluate the significance of enriched reads regions and improves the spatial resolution of binding sites since it combines the information of both sequencing tag position and orientation. There are four main steps in MACS2 analysis: removing redundant reads, adjusting read position, calculation peak enrichment, and estimating the empirical false discovery rate (Figure 11)[28]. The control samples are used in the peak calling to increase the specificity.

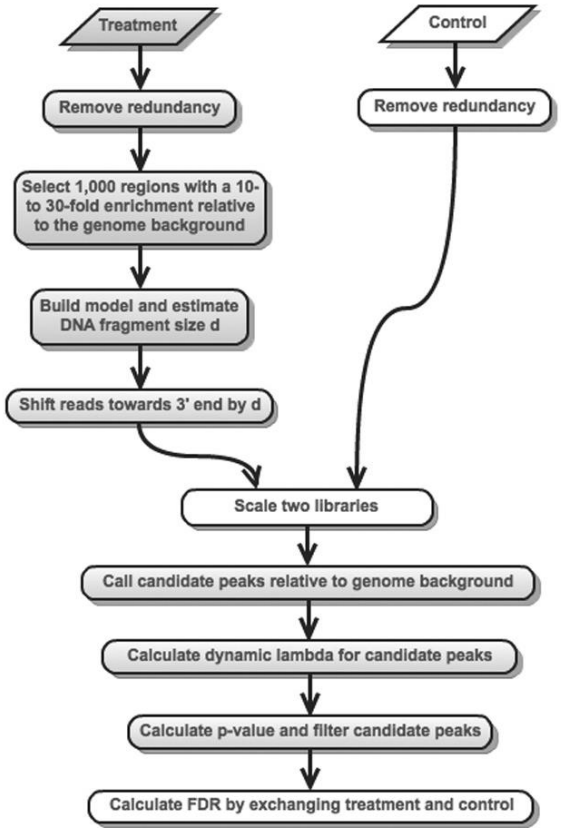


Figure 11: Workflow of MACS2

The first step of MACS2 is removing redundant reads. With a high sequencing coverage, over amplification

of DNA fragments come from ChIP-seq data by polymerase chain reaction (PCR) may cause the same read to be sequenced repeatedly. MACS2 is designed to remove the redundant reads at first based on the parameter set by user without changing the input experiment mapped ChIP-seq reads and control reads to yield more reliable peak calls for downstream analysis. MACS2 will keep no more than one read per genomic location and remove the same one if using default option[28].

The next step is to adjust read position based on fragment size distribution. During the sequencing, the short fragments is equally to sequenced the 5' end of either strands. The reads mapped to the positive and negative strands are likely to appear to position offsetting to the left and right of the real interaction location, resulting in a bimodal enrichment pattern flanking the precise interaction location. In order to extend ChIP-seq reads to represent the original ChIP-DNA fragments, MACS2 first slides a window with a width of about twice the size of the sheared chromatin to identify regions of moderate enrichment and for each peak, MACS2 will calculate the reads from the positive and negative strands separately (Figure 10). All the reads that from the positive and negative strands are extended in the 3' direction until the estimated DNA fragment size d is gained[28].

After Adjusting, MACS2 slides the size 2d window to identify the regions that are significantly enriched in reads across the genome (Figure 12). There are many factors affect the reads enrichment distribution. To minimize the error, MACS2 uses the dynamic parameter λ_{local} to model the number of reads from a genomic region as a Poisson distribution. Considering the local factors, dynamic λ_{local} has been used to vary along the genome instead of a constant value of λ. Based on the Poisson distribution, MACS2 calculates an enrichment p-value for every candidate region and those that lowering the threshold are reported as the final peaks.

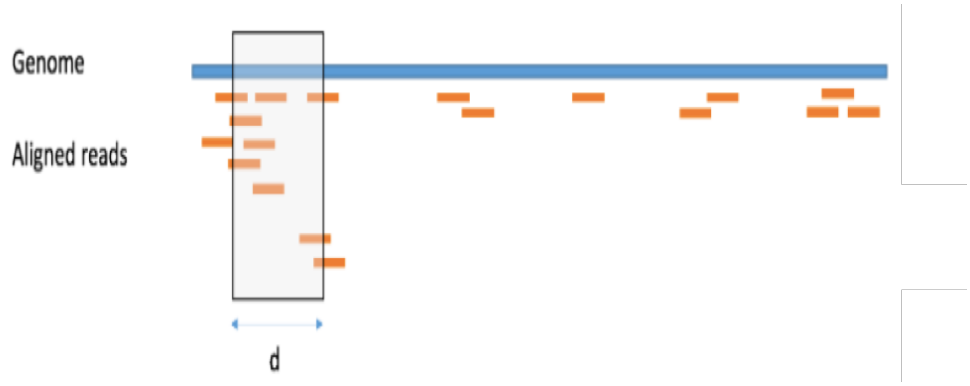


Figure 12: MACS2 slide window to adjust read position based on the fragment size.

Finally, MACS2 uses the control sample to estimate the empirical false discovery rate (FDR) for every peak by exchanging the ChIP-seq and control samples and identifying peaks in the control sample with the same parameters used for the ChIP-seq sample. Any peaks that are identified in the control sample are regarded as false positives.

Before the peak calling, I used DeepTools to generate bedGraph files from the BAM files, which was the required input file format of MACS2 bdgpeakcall function. DeepTools is a suite of python tools developed for efficient analysis of high-throughput sequencing data. The function I used was bamCoverage and set the binSize as 1000. The function bamCoverage takes an alignment of reads as input and generates a coverage track (bigwig or bedGraph) as output. The coverage is calculated as the number of reads per bin. The bedgraph format allows display of continuous-valued data in track format. The data exported in the bedGraph format are preserved in their original state.

In the peak calling step, I used the bdgpeakcall function for narrow peak calling setting the estimated fragments size as 200.

2.5.4 Differential binding analysis of ChIP-seq peak data

DiffBind is a R Bioconductor package that provides options to do some differential binding analysis for ChIP-

seq peak data. DiffBind works primarily with peak regions that associates mapped sequencing reads and the score indicating the strength of the peak together to help downstream analysis.

At first, DiffBind reads in a set of peaksets from the MACS2 and associates metadata. Those peaksets can provide an insight into the potential occupancy of the proteins that identified by ChIP-seq. DiffBind can perform some plotting to determine how the occupancy maps agree with other samples such as between biological replicates or within groups of samples representing a common condition. The output of the occupancy analysis is like a consensus peakset, which represent an overall set of candidate binding sites for further analysis. Then DiffBind uses the sequence reads files to count the number of reads that overlap each peak for each sample. In order to give more standardized peak regions, DiffBind re-centers and trims peaks in the consensus peakset by calculating their summits. The result of counting uses the normalized read counts for each sample at potential protein binding site to form a binding affinity matrix in which the samples can be re-clustered using affinity. The main function of DiffBind is the differential binding affinity analysis. After establishing a contrast that divide the samples into groups to be compared, the binding sites that are statistically significantly differentially bound between sample groups can be identified using the DESeq2 algorithm. Each candidate binding site will be assigned a p-value and false discovery rate to indicate confidence that they are differentially bound. Finally, the DiffBind provide functions to report and plot the results to give an overview of the results of the analysis[29]. I used Diffbind to do differential peak analysis and generate correlation heatmap. Fragment size was set to 300, as appropriate for enhancers, while summit was set to 200. DBA_SCORE_RPKM_FOLD was set as the algorithm to normalize the data.

2.6 Visualization of the expression profiling with the ChIP-seq peaks

After getting the results from the RNA-seq and ChIP-seq result, I used Integrative Genomics Viewer (IGV) to scale the large data sets and flexibly integrate multiple data types together. IGV is a visualization tool for

interactive exploration of large and integrated genomic datasets. It supports integration of a large scale of genomic data types like aligned sequence reads and gene annotation [30].

I used the IGV to visualize the integration of ChIP-seq peak files from two samples, gene expression profile that performed by RNA-seq and mouse enhancers and promoters identified in multiple tissues from ENCODE database together to give an epigenetic profiling of active enhancers in mouse retinal ganglion cells and compare them with the known enhancers and promoters. The known enhancer and promoter data were from the 19 types of tissues and cells. The same experiment was applied to a diverse set of tissues and cell types in the mouse to produce a map of about 300000 murine of cis-regulatory sequences[31]. The ENCODE database define tissue-specific enhancers and promoters and I merged the data of enhancers and promoters respectively.

2.7 Motif analysis

2.7.1 Motif discovery

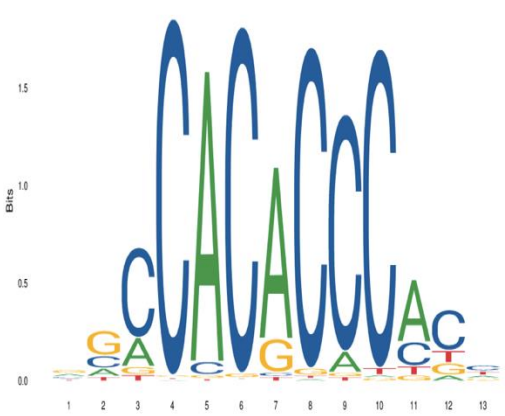
A motif is a sequence pattern that occurs in a group of related sequences repeatedly. To predict which transcription factors that are expressed in retinal ganglion cells can bind to the active enhancers, I used HOMER software to annotate the peaks and find sequence motifs through them. Hypergeometric Optimization of Motif Enrichment (HOMER) is a suite of tools for motif discovery and next-generation sequencing data analysis. HOMER can be used as a *de novo* motif discovery algorithm to find 8-12 base pair motifs in a large genomics data. I used HOMER *findMotifsGenome* to analyze genomic positions for enriched motifs through ChIP-seq peak regions. This will perform *de novo* motif discovery and also check the enrichment of known motifs[32]. The input files of the motif discovery analysis is the BED file that contain the information of ChIP-seq peaks such as chromosome, start & end position and the type of strand. The - size parameter was set as 400 to tell HOMER what the size of the ChIP-seq peaks. The HOMER will first verify the peak file to make sure that it is contained valid and unique peaks without replicates. Then sequences are extracted from the genome corresponding to the

regions of the peaks. HOMER select background regions as a control for motif discovery randomly to support a reliable result. There is a unique procedure that is provided by HOMER. The sequence would be normalized to remove bias introduced by lower-order oligo sequences. Finally, HOMER looks for enriched motifs in the peak regions with a n assigned p-value by screening the library of reliable motifs against the target and background sequences and will give the motifs of length 8, 10, 12. The output of HOMER will provide the enriched motifs files from the *de novo* motif discovery, separated by motif length.

2.7.2 Motif comparison

After finding the enriched motifs through peak regions, I used MEME suite tools to cross-reference the enriched motifs to the transcription factors. The MEME suite are motif-based sequence analysis tools that help us to use the enriched motifs for further analysis[33]. Tomtom is one of the motif comparison tool in MEME suit. We can use Tomtom to compare one or more motifs against a database of known motifs to rank the motifs that find in the database and provide an alignment for each significant match between input motifs and known motifs database. When given query motifs, Tomtom searches them against the database of target motifs and provide reports for each query a list of target motifs that matched, ranked by p-value[34]. I chose JASPAR CORE 2018 as target database. JASPAR is the largest open-access database of curated and non-redundant transcription factor (TF) binding profiles for TFs across multiple species from six different taxonomic groups. The JASPAR CORE data are derived from published and experimentally defined transcription factor binding sites for eukaryotes. A PFM summarizes the DNA sequences that an individual TF bind to by counting the number of occurrences of each nucleotide at each position within the TF-binding sites. The transcription factor binding profiles are stored as position frequency matrices (PFMs) and transcription factor flexible models (TFFMs). PFMs can be also converted to probabilistic models - position weight matrices (PWMs), which can be used to predict TF-binding sites in DNA sequences[35]. Figure 13 shows an example of the motif logo and position

frequency matrices of transcription factor klf9 which is highly expressed in the mouse retinal ganglion cells identified by our RNA-seq data[35]. By using the JASPAR database, query motifs can be compared with the database that contains individual TF-bind sites and the reports of the Tomtom can provide the specific potential transcription factors that can bind to the target motifs.




```
ALPHABET= ACGT

strands: + -

Background letter frequencies
A 0.25 C 0.25 G 0.25 T 0.25

MOTIF MA1107.1 KLF9
letter-probability matrix: alength= 4 w= 13 nsites= 5754 E= 0
 0.306917  0.203163  0.328815  0.161105
 0.165450  0.251825  0.494786  0.087939
 0.215676  0.675356  0.061522  0.047445
 0.004171  0.983490  0.006778  0.005561
 0.937261  0.042579  0.012513  0.007647
 0.004171  0.977233  0.013730  0.004866
 0.801877  0.020507  0.157108  0.020507
 0.007125  0.961244  0.021550  0.010080
 0.079944  0.887904  0.016858  0.015294
 0.009559  0.960028  0.009211  0.021203
 0.612096  0.238443  0.058220  0.091241
 0.082030  0.571081  0.171011  0.175878
 0.210115  0.382343  0.150156  0.257386
```

Figure 13 The motif logo and position frequency matrices.

2.8 Promoter motif analysis

Motif discovery and comparison are also used in finding motifs in promoter regions of RGCs. All the genes that are highly expressed in RGCs (threshold is FPKM > 1) are used to perform motif analysis of their promoters. Promoter regions are found by the EPDnew promoter database. EPDnew is a collection of databases of experimentally validated promoters for selected model organisms[36]. EPDnew mouse database contains 21239 promoters which can be searched with the select tool by the EPDnew ID, ENSEMBL gene ID or RefSeq ID. I selected 400 base pairs as the length of promoter regions and performed the motif analysis including the motif discovery and motif comparison. Each motif is also associated the TFs that have the potential to bind to it.

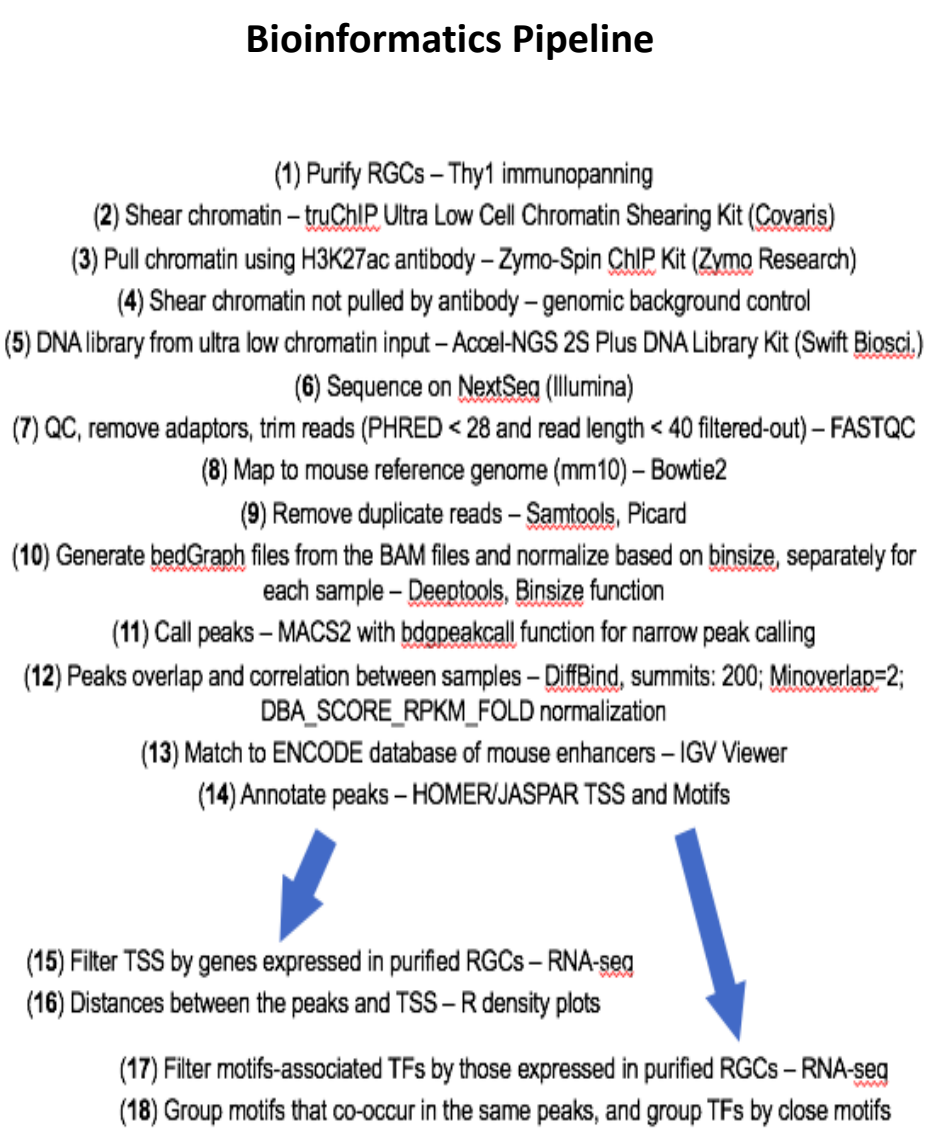


Figure 14 Bioinformatics pipeline

Chapter 3: Analysis and results

3.1 RNA-seq data

The Trimmomatic dropped 18818507 (13.40%) of total 140441332 reads in embryonic day 18 polyA RNA-

seq data and dropped 17116094 (12.76%) of total 134119404 reads in postnatal day 5 data using the same prefix pairs. According to the Histat2 reports, the over alignment rate of embryonic day 18 data is 99.78% and 99.80% for postnatal data (Figure 15). After getting the BAM files from both ages, I merged them together to get a better view in the IGV so that I could find the highly expressed transcripts throughout the whole genome.

```
Embryonic day 18
82830505 reads; of these:
  82830505 (100.00%) were paired; of these:
    474113 (0.57%) aligned concordantly 0 times
    73488669 (88.72%) aligned concordantly exactly 1 time
    8867723 (10.71%) aligned concordantly >1 times
    ----
    474113 pairs aligned concordantly 0 times; of these:
      218644 (46.12%) aligned discordantly 1 time
    ----
    255469 pairs aligned 0 times concordantly or discordantly; of these:
      510938 mates make up the pairs; of these:
        364739 (71.39%) aligned 0 times
        74140 (14.51%) aligned exactly 1 time
        72059 (14.10%) aligned >1 times
99.78% overall alignment rate
```

```
Postnatal day 5
89330175 reads; of these:
  89330175 (100.00%) were paired; of these:
    523026 (0.59%) aligned concordantly 0 times
    81471795 (91.20%) aligned concordantly exactly 1 time
    7335354 (8.21%) aligned concordantly >1 times
    ----
    523026 pairs aligned concordantly 0 times; of these:
      273238 (52.24%) aligned discordantly 1 time
    ----
    249788 pairs aligned 0 times concordantly or discordantly; of these:
      499576 mates make up the pairs; of these:
        364793 (73.02%) aligned 0 times
        63799 (12.77%) aligned exactly 1 time
        70984 (14.21%) aligned >1 times
99.80% overall alignment rate
```

Figure 15: The Histat2 mapping reads result of embryonic day 18 and postnatal day 5.

The GTF files that were generated by Cufflinks could give us the FPKM value of each gene which indicated the expression level of it.

3.2 ChIP-seq data

Each sample that was used as the two biological replicates included RGCs purified from multiple mice of various ages (embryonic day 18, postnatal day 5, and adult). The data sources of each ChIP-seq sample are showed in Figure 16.

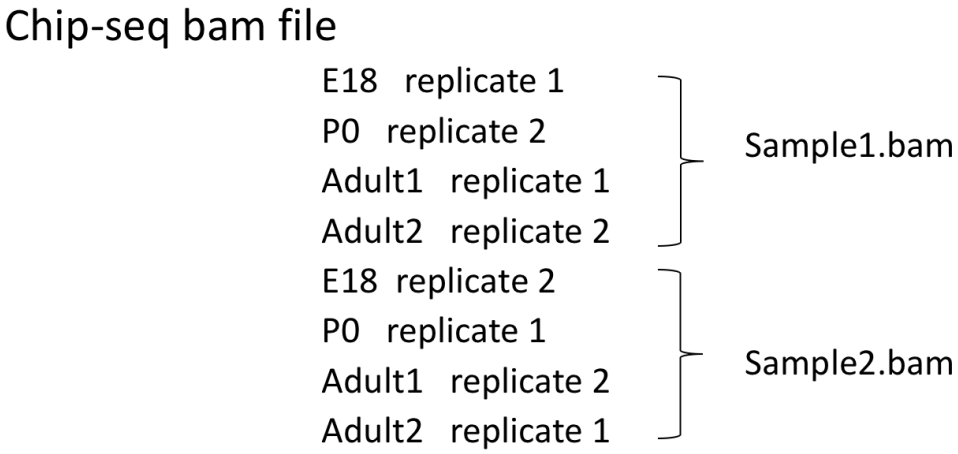


Figure 16: The sources of each ChIP-seq sample BAM file.

The Figure 17 shows the the probability density of the reads counts distribution in the genome for the two biological replicate samples. The blue and red line represent the reads counts of sample1 and sample2 respectively. The y-axis represents the mass of read counts per bin which normalized to millions of bins. The figure was plotted by the R package Rsurbread, using Kernel density estimates (KDE). The counts of each sample were generated by the featureCounts function in Rsurbread. The density distribution was plotted by the dens() function in R.

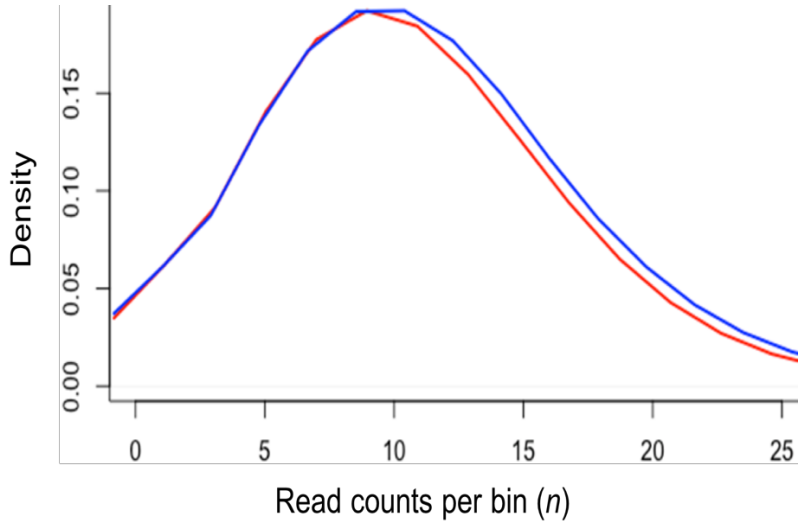


Figure 17 Probability density of the read counts distribution in the genome for the two biological replicate samples (blue and red lines).

629 peaks that co-enriched in two samples were statistically significant identified by Diffbind through MACS2 identifying peaks. The result of differential peaks analysis is displayed in Figure 18 (Pearson correlation coefficients *r* are shown).

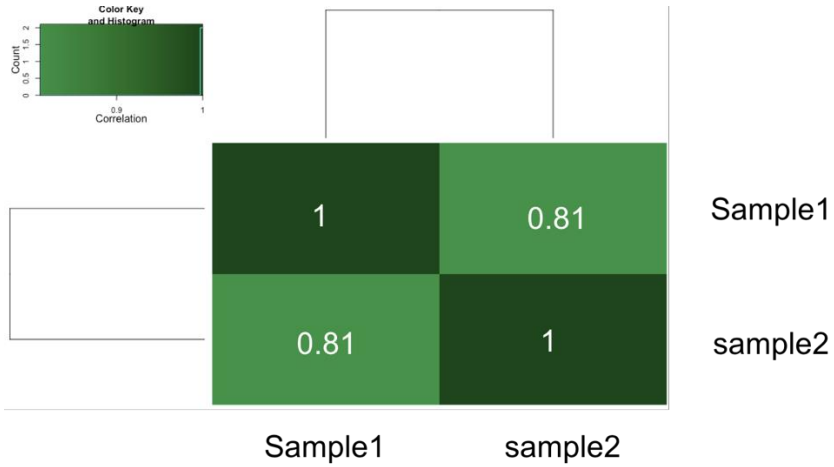


Figure 18 Correlation heatmap of sample1 and sample2 by Diffbind.

After getting the co-enriched peaks, I visualized the active enhancers detected in the RGCs (ChIP-seq peaks alignment to the mouse reference genome mm10) in the IGV Viewer and cross-referenced them to the mouse enhancers and promoters identified in multiple tissues or cell type from the ENCODE database (Figure 19). The tissue-specific enhancers are from bone marrow, cerebellum, cortex, heart, limb, liver, intestine, kidney, lung, mouse embryonic fibroblast, embryonic stem cells, olfactory bulb, placenta, spleen, testis and thymus. I discovered a number of novel enhancers that maybe only specific to retinal ganglion cells because we used highly purified retinal ganglion cells which only comprise 1% of all retinal cells.

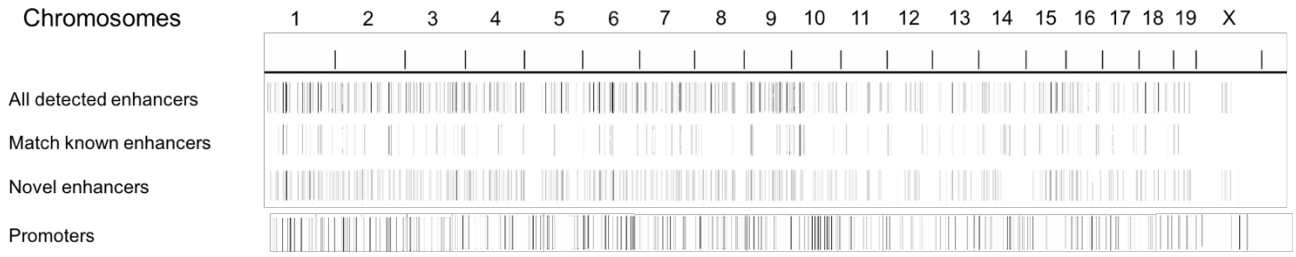


Figure 19 IGV Viewer visualization of peaks alignment to the mouse genome (mm10) and cross-reference to the ENCODE database of mouse enhancers and promoters.

The scatter plot (Figure 20) indicates correlation between concentration of peaks that are co-enriched in biological replicate samples. The concentration value of each sample was derived from normalized read counts of the 629 peaks co-enriched in each sample. Read counts normalized by FPKM are shown as log2. Concentration values were generated with Diffbind using the dba.report() function. Peak overlap parameters were set as minOverlap = 2, summits = 200 and peaks normalization function was set as DBA_SCORE_RPKM_FOLD. I used ggplot2 to generate the scatter plot. The Pearson correlation coefficient r is 0.9 ($p < 0.001$, two-tailed). Density of the overlapping data points was color-coded with the scale bar of color on the right.

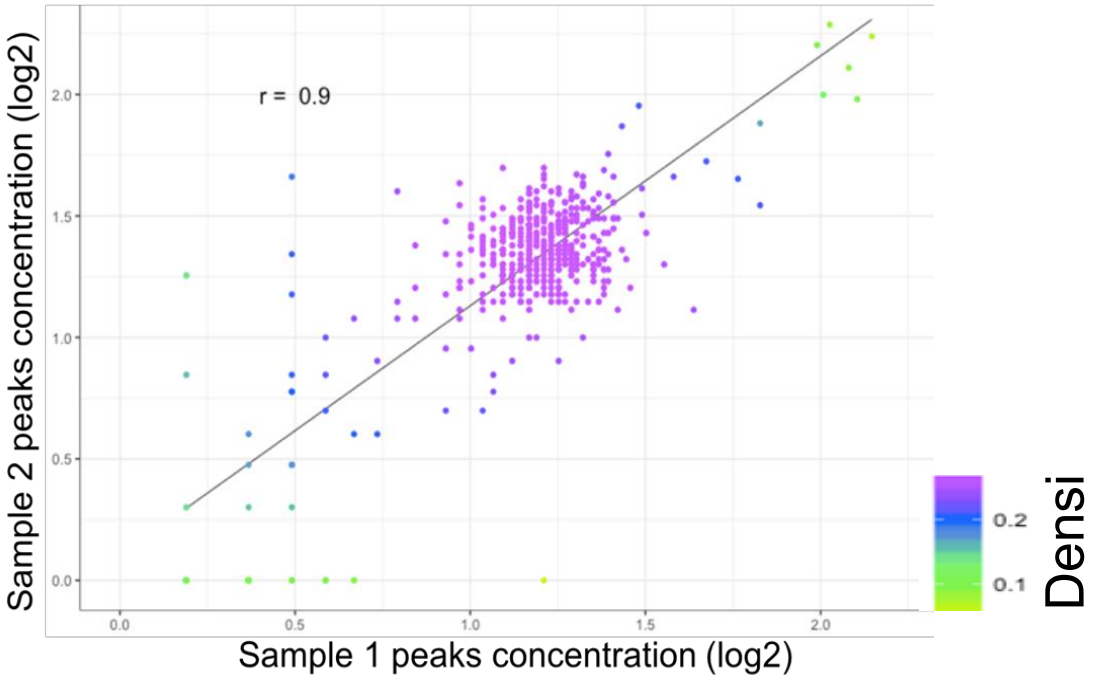


Figure 20 Scatter-plot of the peaks concentration of sample1 and sample2.

The active enhancer peaks were annotated by HOMER. TSSs from genes expressed in RGCs at least 1 FPKM, as determined by RNA-seq profiling of purified RGCs from multiple mice of various ages, were used for computing the distances. TSSs and entries for non-expressed or the value of FPKM<1 were filtered out from the GTF file that used in HOMER for peaks annotation and calculations of distances from TSSs. Then probability density of peaks distances from the nearest downstream TSS that is active in RGCs was plotted with R package Rsurbread, using Kernel density estimates (KDE). The density distribution was plotted using the dens() function. Y-axis shows the mass of peaks per kb log10 (Figure 21).

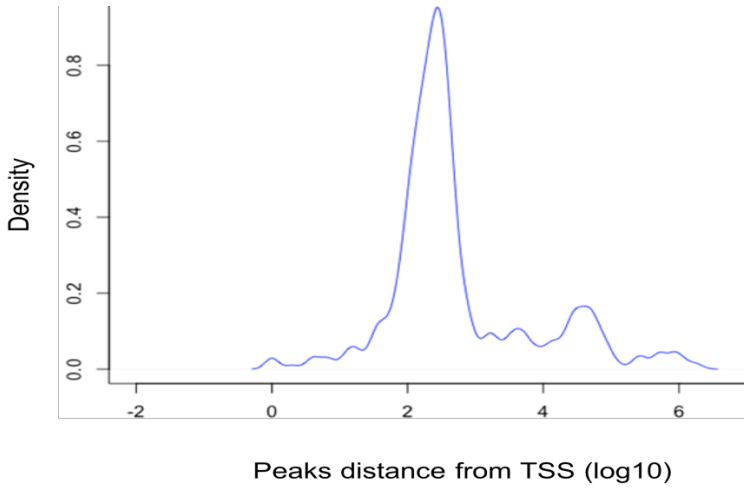


Figure 21: Probability density of peaks distances from the nearest TSS that is active in RGCs.

Figure 22 displays the distribution of peaks' annotated genes ranked by expression percentiles after filtering out non-expressed genes in RGCs. The genes that annotated genes were ranked by their expression percentiles (100% = max expression). The distributions of genes are shown before (red line) and after (blue line) filtering out non-expressed or FPKM < 1 in RGCs genes. The values of FPKM were determined by RNA-seq profiling of purified RGCs from multiple mice of various ages. The distribution of genes' expression percentiles was plotted with R-package Rsubread, using Kernel density estimates (KDE). Density distribution was plotted by the den() function. Mass of genes per percentile are showed on y-axis.

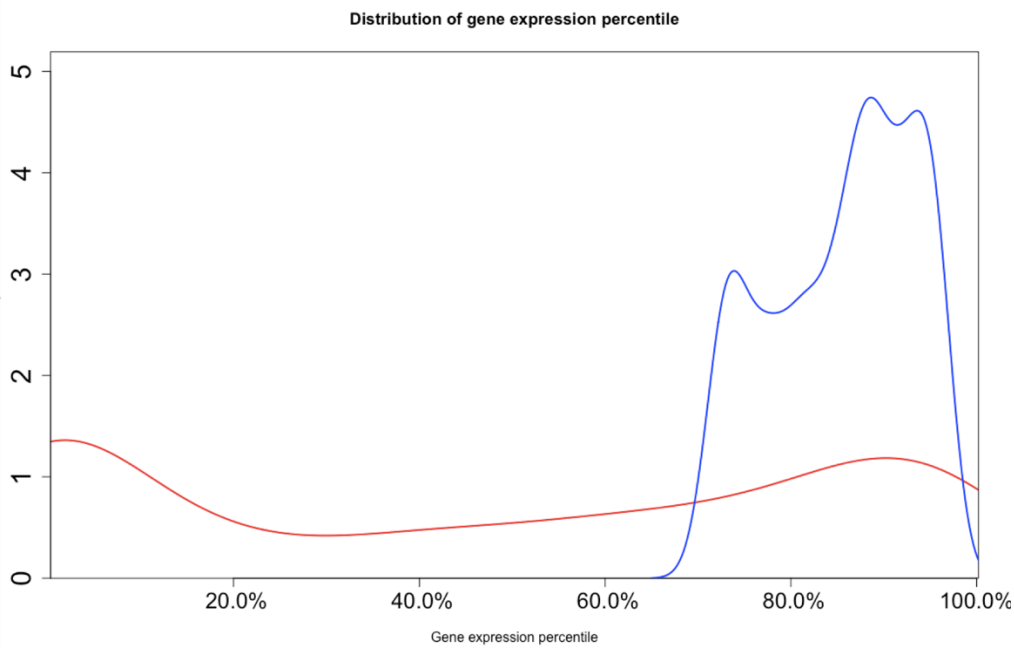


Figure 22: Probability density of peaks annotated by their expression percentile (100% = max expression)

A bar plot was generated to show the average expression of genes (two-tailed; error bars, 1 SEM). As the bar plotted shown (Figure 23), the average expression of peaks annotated genes is significantly higher after filtering, demonstrating improved predictive power compared to default unfiltered HOMER peak annotation based on all mouse genes. The p value that was tested by t-test is less than 0.0001.

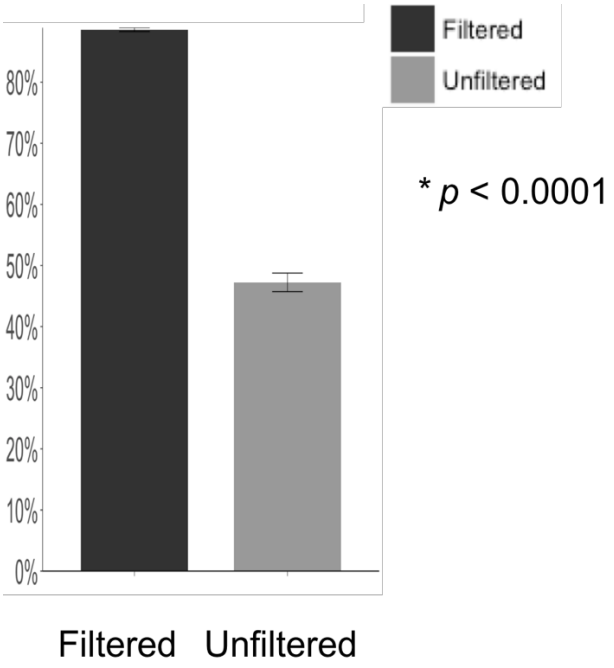


Figure 23: A bar plot that shows the average expression of genes (two-tailed; error bars, 1 SEM).

By motif discovery tool HOMER, I identified 74 enriched motifs through active enhancers that were found by ChIP-seq data in RGCs. After comparing to the JASPAR database, transcription factors filtered by the expression level in RGCs were associated to each motif, as show in Table 1. Each motif is counted through the peaks and given the result of the number of times occur in peaks, the number of peaks in which this motif is found at least once, the number of TFs that have the potential to bind to it and the FPKM value of the TFs that express in RGCs.

3.3 Motif analysis

I next focused on the distribution of enriched motifs and Transcription factors(TFs) that are expressed in RGCs by peaks. The pie graph (Figure 24) indicates the distribution of transcription factors that are expressed in RGCs. The transcription factors are classified into three types. The first one is transcription factors that their transcription factor-binding sites are associated with motifs found in the identified peaks, while the second

type is transcription fators that their transcription factor-binding sites are associated with motifs that are not found in the identified peaks. The last type of transcription factors are those their transcription factor-binding sites are unknown to data. 32% of the transcription factors that are expressed in RGCs can be found enriched motifs through identified peaks in RGCs.

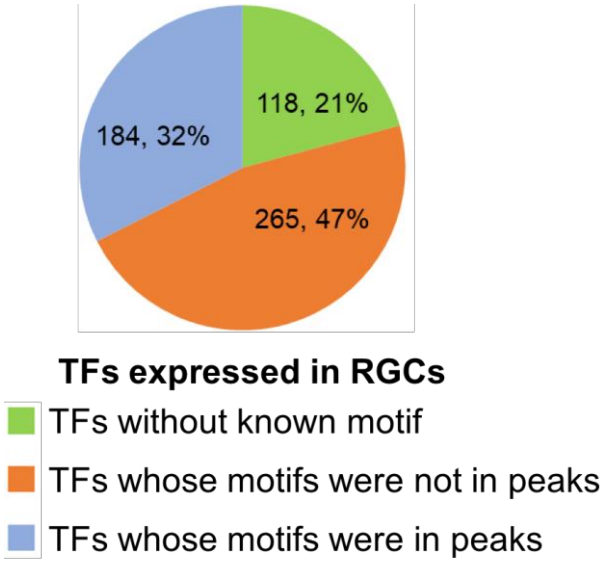


Figure 24: Distribution of TFs that are expressed in RGCs.

Probability density plot of identified motifs and transcription factors per peak is shown in Figure 25. The red line is the probability density line of enriched motifs per identified peaks. The blue line displays the density distribution of highly expressed transcription factors per identified peaks based on known associations with the identified motifs. KDE number distribution and density plot of R package are used in the figure. Y-axis shows the mass of numbers per peak.

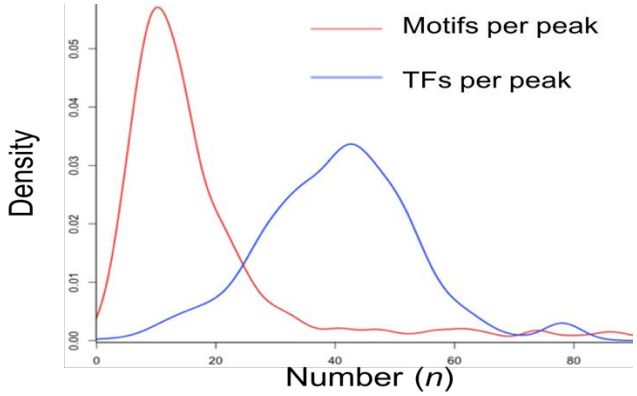


Figure 25: Probability density of identified motifs and TFs per peak.

Figure 26 and Figure 27 demonstrate the distribution of transcription factors that have associations with the enriched motifs and the distribution of enriched motifs through identified enhancers and promoters. The total number of transcription factors that are found in the enhancers and promoters is 277. 219 of them are found from the enhancers with the proportion of 79%. 119 of the transcription factors are found within the enriched motifs from the promoter, while 61 of them are overlapped by the transcription factors in enhancers. As for the distribution of enriched motifs, 81.4% of total motifs are found through the enhancers, while 55% of the motifs can be found in the promoters.

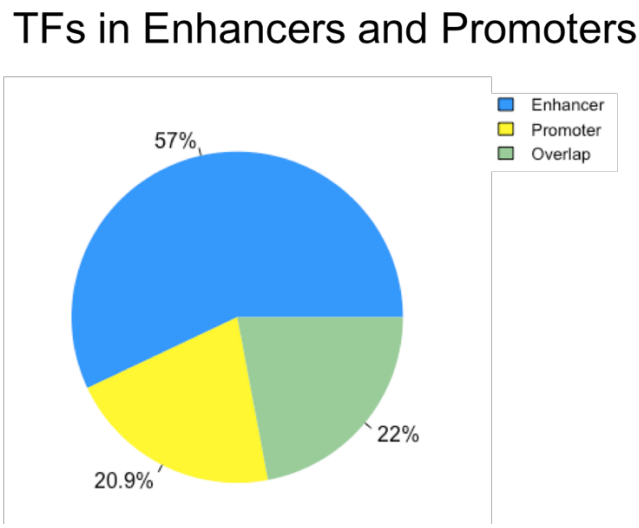


Figure 26: Distribution of TFs that have associations with the enriched motif through identified enhancers and promoters.

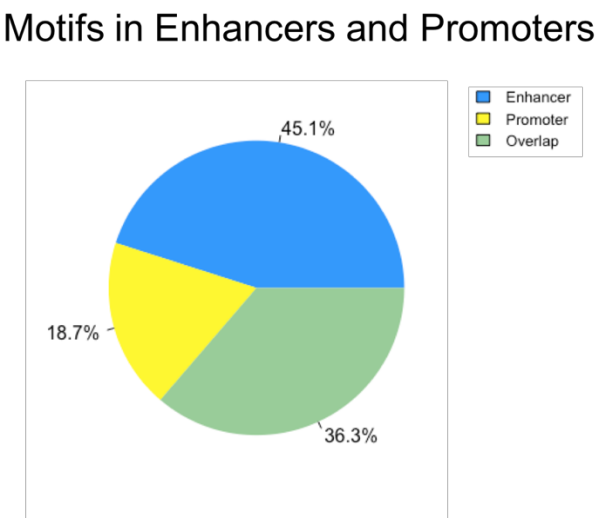


Figure 27: Distribution of enriched motifs through identified enhancers and promoters.

Two bar plots were generated to show the occurrence of motifs in enhancer peaks and promoter peaks respectively. The bar plot of Figure 28 shows the enhancer motifs occurrence, while the Figure 29 displays the promoter motifs occurrence. The occurrence of promoter motifs is higher than the occurrence of enhancer motifs. The p value that was tested by t-test is less than 0.0001in both situations.

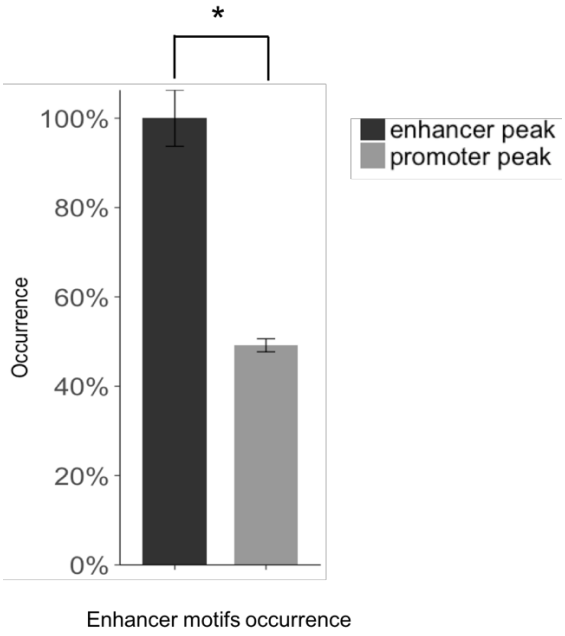


Figure 28: The bar plot of enhancer motif occurrence within enhancer peaks and promoter peaks (two-tailed; error bars, 1 SEM).

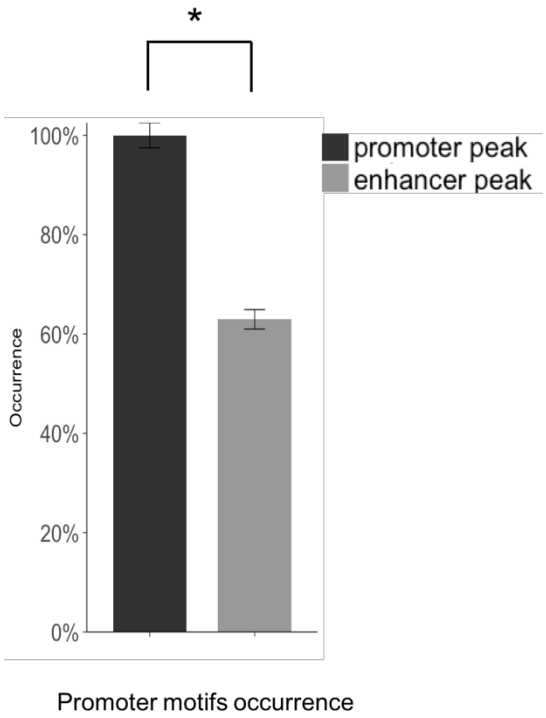


Figure 29: The bar plot of promoter motifs occurrence within promoter peaks and enhancer peaks

(two-tailed; error bars, 1 SEM).

Figure 30 shows the position of the ChIP-seq peak in biological replicates (the first and the second tracks), and the first downstream gene, Rab10, that is expressed in RGCs, as shown in the track below. Last track shows positions of the known mouse enhancers from multiple cell types, including one that matched the enhancer identified in RGCs.

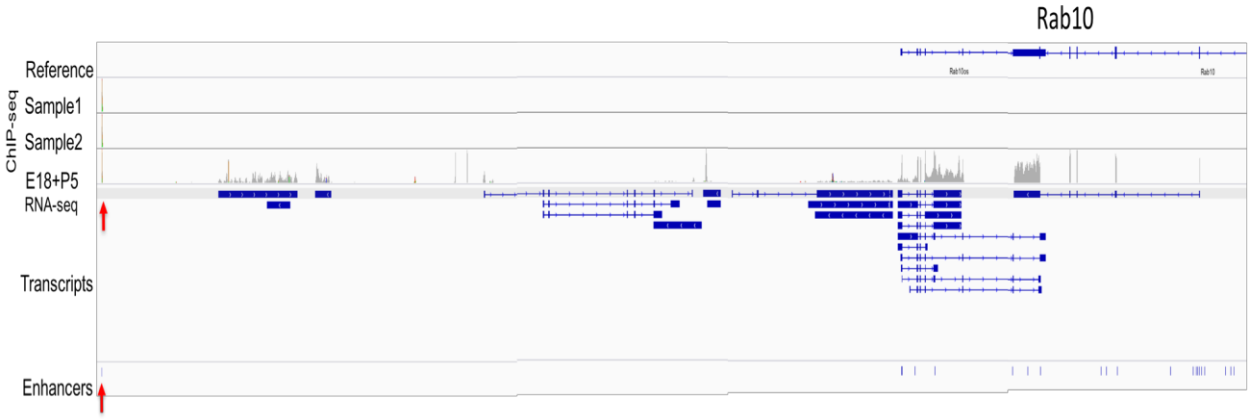


Figure 30: Example 1 of enhancers and downstream genes

Figure 31 shows the position of the ChIP-seq peak in biological replicates (the first and the second tracks), and the first downstream gene, Dpp9, as well as following genes Mir7b, Fem1a, Uhrf1 and kdm4b, which are all highly expressed in RGCs, as shown in the track below. Last track shows positions of the known mouse enhancers from multiple cell types, without a match to the novel enhancer that is identified in RGCs.

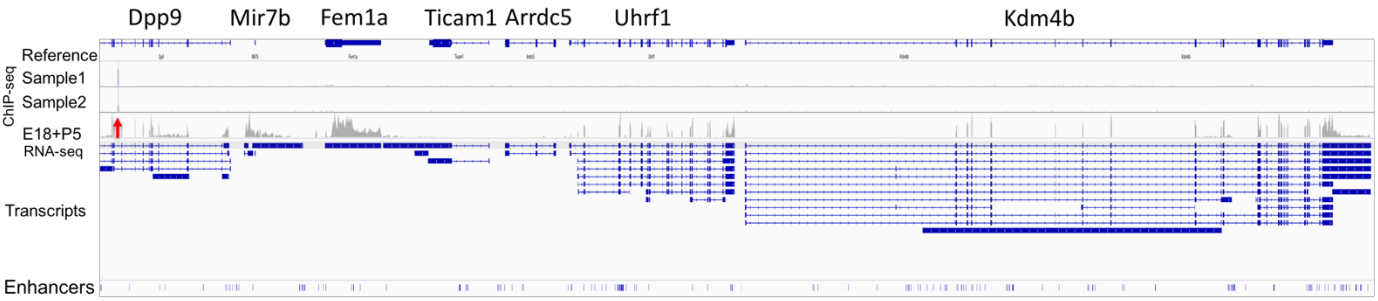


Figure 31: Example 2 of enhancers and downstream genes

Figure 32 shows the position of the ChIP-seq peak in biological replicates (the first and the second tracks), and the first downstream gene, Pla2r1, which is expressed in RGCs, as shown in the track below. Enlarged peak is shown below. The peak is a novel enhancer, as there were no matches in the database of known enhancers. Positions of known motifs within the enhancer, as predicted by HOMER, are indicated (5 of 6 motifs are unique). Two groups of3 motifs each are within sufficient proximity (within each group) for the TFs that bind to them to form transcriptional complex. TFs expressed in RGCs with TF-binding sites within these motifs are shown under each motifs, along with their corresponding TF-binding sites nucleotide consensus sequences logos, respectively.

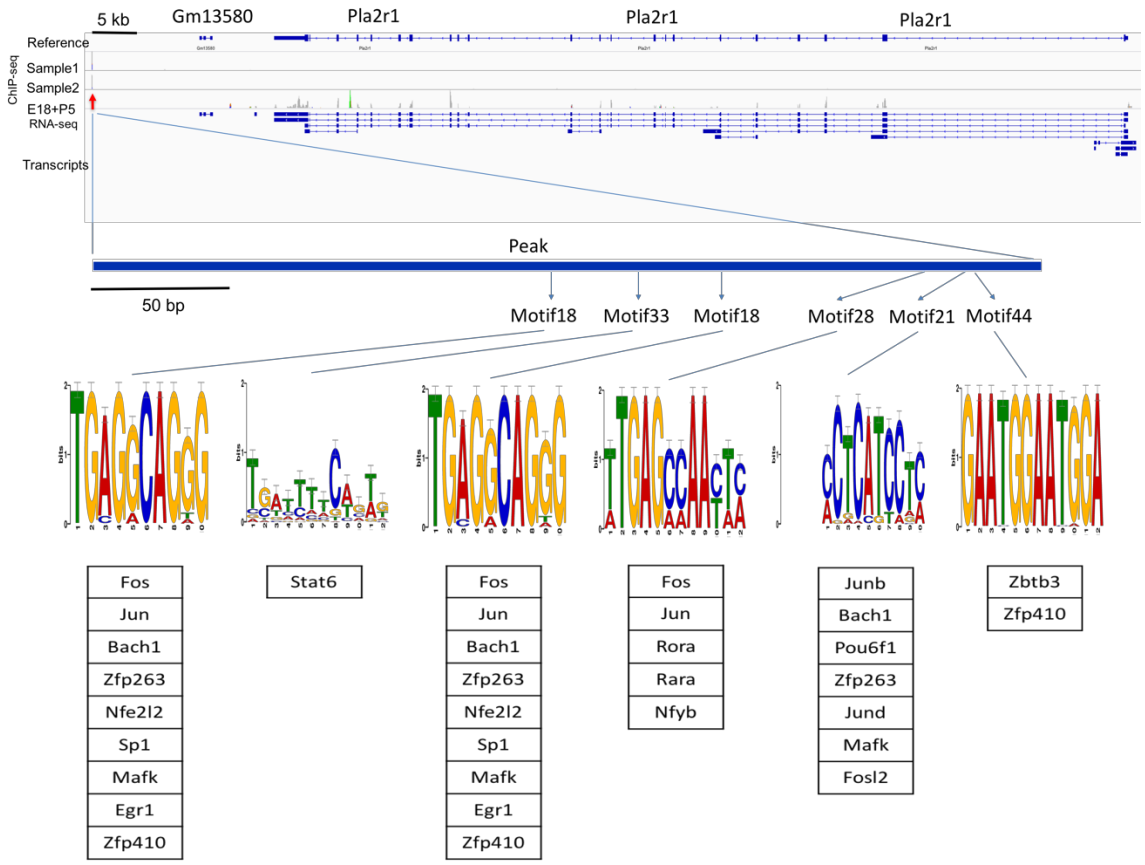


Figure 32: Motifs and TFs associated with enhancers

Chapter 4: Discussion

I integrate the mRNA-seq and ChIP-seq data to generate the epigenetic H3K27ac profiling of active enhancers in purified mouse RGCs and predicte 629 active enhancers through the genome. After being compared to the enhancers identified in multiple cell types from ENCODE database, I discovered that about half of the enhancers active in RGCs are known from other mouse cell types, and other half are putative novel enhancers that may be unique to RGCs and other, potentially also rare, cell types.

The 629 peaks are annotated by HOMER to reference to the nearest gene that has the potential to be regulated by the corresponding enhancer. Accuracy of peak annotations using HOMER is significantly improved if GTF file is filtered to exclude non-expressed genes.

TF-binding sites motif analysis in the peaks identified 71 known motifs and predicted 3 novel motifs, which may be unique to RGCs. Each known motif has one or more TFs expressing in RGCs that have potential to bind to it. 184 of TFs expressed in RGCs matched the TF-binding sites motifs in the peaks.

TF-binding sites proximity analysis in the peaks predicted the TFs that could form transcriptional complexes to regulate gene expression in RGCs.

References

1. Masland, R. H. Neuronal diversity in the retina. *Current Opinion in Neurobiology* **11,** 431–436 (2001).
2. Hoon, M., Okawa, H., Della Santina, L. & Wong, R. O. L. Functional architecture of the retina: Development and disease. *Progress in Retinal and Eye Research* **42,** 44–84 (2014).
3. Jiang, Y. *et al.* Transcription factors SOX4 and SOX11 function redundantly to regulate the development of mouse retinal ganglion cells. *J. Biol. Chem.* **288,** 18429–18438 (2013).
4. Baden, T. *et al.* The functional diversity of retinal ganglion cells in the mouse. *Nature* **529,** 345–350 (2016).
5. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* (2011). doi:10.1038/nrg3095
6. Nguyen, T. A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26,** 1023–1033 (2016).
7. Danino, Y. M., Even, D., Ideses, D. & Juven-Gershon, T. The core promoter: At the heart of gene expression. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1849,** 1116–1131 (2015).
8. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: Five essential questions. *Nature Reviews Genetics* **14,** 288–295 (2013).
9. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.* **107,** 21931–21936 (2010).
10. Tomovic, A., Stadler, M. & Oakeley, E. J. Transcription factor site dependencies in human, mouse and rat genomes. *BMC Bioinformatics* **10,** 339 (2009).
11. Reis-Filho, J. S. Next-generation sequencing. *Breast Cancer Res.* **11 Suppl 3,** S12 (2009).
12. Illumina. Illumina sequencing technology. *Technol. Spotlight Illumina Seq.* 1–5 (2010). doi:10.1016/S0167-7799(03)00189-6
13. Adams, J. Transcriptome: connecting the Genome to Gene function. *Nat. Educ.* **1,** 1–3 (2008).
14. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63 (2009).
15. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15,** (2014).
16. Furey, T. S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics* **13,** 840–852 (2012).
17. Ebbert, M. T. W. *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* **17,** (2016).
18. Trakhtenberg, E. F. *et al.* Cell types differ in global coordination of splicing and proportion of highly expressed genes. *Sci. Rep.* **6,** (2016).
19. Zhang, Y. *et al.* An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34,** 11929–11947 (2014).
20. Paralkar, V. R. *et al.* Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. *Blood* **123,** 1927–1937 (2014).
21. Kurimoto, K. *et al.* Quantitative dynamics of chromatin remodeling during germ cell

specification from mouse embryonic stem cells. *Cell Stem Cell* **16,** 517–532 (2015).
22. Sundaram, A. Y. M. *et al.* A comparative study of ChIP-seq sequencing library preparation methods. *BMC Genomics* **17,** (2016).
23. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
24. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578 (2012).
25. Shin, H., Liu, T., Duan, X., Zhang, Y. & Liu, X. S. Computational methodology for ChIP-seq analysis. *Quantitative Biology* **1,** 54–70 (2013).
26. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
27. Leleu, M., Lefebvre, G. & Rougemont, J. Processing and analyzing ChIP-seq data: From short reads to regulatory interactions. *Brief. Funct. Genomics* **9,** 466–476 (2010).
28. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7,** 1728–1740 (2012).
29. Stark, R. & Brown, G. DiffBind : differential binding analysis of ChIP-Seq peak data. *Cancer Res.* 1–27 (2011).
30. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26 (2011).
31. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).
32. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).
33. Bailey, T. L. & Elkan, C. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. *Proc. Second Int. Conf. Intell. Syst. Mol. Biol.* 28–36 (1994). doi:citeulike-article-id:878292
34. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8,** (2007).
35. Khan, A. *et al.* JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46,** D260–D266 (2018).
36. Dreos, R., Ambrosini, G., Périer, R. C. & Bucher, P. The Eukaryotic Promoter Database: Expansion of EPDNew and new promoter analysis tools. *Nucleic Acids Res.* **43,** D92–D96 (2015).